

EXHIBIT 7

**IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF PENNSYLVANIA**

IN RE: DIISOCYANATES ANTITRUST
LITIGATION

This Document Relates to All Actions

Master Docket Misc. No. 18-1001

MDL No. 2862

**DECLARATION OF MAURA R. GROSSMAN IN SUPPORT OF PLAINTIFFS'
MOTION TO COMPEL DEFENDANTS TO USE CERTAIN SEARCH TERMS AND
TAR METHODOLOGIES**

Pursuant to 28 U.S.C. § 1746, I, Maura R. Grossman, declare as follows:

1. I am a Research (Full) Professor in the David R. Cheriton School of Computer Science at the University of Waterloo, an Adjunct Professor of Law at Osgoode Hall Law School of York University, and an Affiliate Faculty Member at the Vector Institute of Artificial Intelligence, all in Ontario, Canada. I also am an eDiscovery attorney and consultant in Buffalo, New York. I am an approved Electronic Discovery Special Master and Mediator for the Western District of Pennsylvania; since 2011, I have served as an eDiscovery special master in 14 high-profile federal and state court cases across the United States. Since 2010, I have taught more than a dozen courses on electronic discovery at Columbia Law School, Georgetown University Law Center, Pace University Law School, and Rutgers Law School-Newark. A copy of my *curriculum vitae* is attached hereto as Exhibit A.

2. I received my Juris Doctorate (J.D.), *magna cum laude*, *Order of the Coif*, from the Georgetown University Law Center. I also hold masters (M.A.) and doctoral (Ph.D.) degrees from The Derner Institute of Advanced Psychological Studies at Adelphi University, and a Bachelor of Arts (A.B.), *magna cum laude*, with Honors in Psychology, from Brown University.

3. Before opening my own law and consulting practice in June, 2016, for 17 years I was a litigator at the New York law firm of Wachtell, Lipton, Rosen & Katz; from 2007 until my departure from the firm, I oversaw the firm's eDiscovery processes. For the past 14 years, my practice has focused exclusively on legal, technical, and strategic issues involving eDiscovery and related matters. I am widely recognized as one of the foremost experts in the field of technology-assisted review ("TAR") technologies and validation processes; in *Rio Tinto v. Vale*, No. 14 Civ. 3042 (RMB) (AJP), 2015 WL 4367250 (S.D.N.Y. July 15, 2015), a case in which Magistrate Judge Andrew J. Peck appointed me as a special master to resolve disputes that arose in connection with the use and validation of TAR, he referred to me as "one of the most knowledgeable (if not *the* most knowledgeable lawyer) about TAR. . . ." (emphasis in original).

Id. at *1.

4. Since 2010, I have personally been involved in hundreds of TAR reviews (the vast majority of which have involved "TAR 2.0," the TAR methodology proposed to be used in this matter). I hold ten patents and two trademarks related to TAR, and have published approximately 30 peer-reviewed papers on TAR and its validation, in addition to approximately 60 other articles and book chapters.

5. My research on TAR with University of Waterloo Computer Science Professor Gordon V. Cormack has been cited approvingly in numerous federal cases, beginning with *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 190 (S.D.N.Y. 2012), the first case to recognize and approve the use of TAR in civil litigation. Notably, our study, *Technology-Assisted Review in e-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, published in volume 17 of the Richmond Journal of Law and Technology (2011) ("Richmond Journal study") has been widely cited in case law, both in the U.S. and abroad. My TAR

research and scholarship is funded by nine grants and fellowships, including from the Ontario Research Fund, the Canadian Foundation for Innovation, and the National Sciences and Engineering Council of Canada.

6. I have provided many eDiscovery training programs on TAR and other eDiscovery-related topics to federal and state court judges, by invitation of the court, and have testified, on several occasions, before the Advisory Committees on the Federal Rules of Civil Procedure and the Evidence Rules, at their invitation.

7. I have been retained by Plaintiffs in the above-captioned matter to consult on discovery issues related to electronically stored information, particularly as relates to search methodologies, TAR, and the validation of both.

8. I submit this declaration in support of Plaintiffs' Motion to Compel Regarding Defendants' Search Terms and TAR Methodologies. I am generally familiar with these proceedings, the issues that have been raised in the Motion, have personal knowledge of the matters set forth herein, and if called upon and sworn as a witness, I could competently testify thereto.

9. Professor Cormack and I first coined the term "technology-assisted review" or "TAR" in the Richmond Journal study (referenced in ¶5 above). We subsequently defined the term more formally in *The Grossman-Cormack Glossary of Technology-Assisted Review*, published in volume 7 of *Federal Courts Law Review* (2013). Basically, TAR is a process for ranking—from most to least likely to be responsive—or for classifying—as responsive or nonresponsive—a document collection, using computer software that learns to distinguish between responsive and non-responsive documents based on coding decisions made by one or more knowledgeable reviewers on a subset of the document collection. The software then

applies what it has learned to the remaining documents in the collection. While TAR tools differ in their specific mechanisms and workflows, all TAR tools either rank or classify documents based on their likelihood of being responsive, and their performance is evaluated through a post-hoc validation process.

10. Search and review efforts are typically evaluated using two metrics drawn from the science of information retrieval. One is referred to as *recall*, which is a measure of *completeness*, reflected by the proportion (*i.e.*, percent) of responsive documents in a collection that have been found through a search or review process, out of all possible responsive documents in the collection. For example, if there were 100 responsive documents in a collection of 1,000 documents, and a search or review effort identified 225 of them for production, 90 of which were actually responsive—assuming, some non-responsive attachments were also produced with the 90 responsive documents—recall would be 90/100, or 90%. The other metric is referred to as *precision*, which is a measure of *accuracy*, or the proportion (*i.e.*, percent) of the documents identified by a search or review process that are actually responsive. Using the same example, precision would be 90/250, or 36%. High recall suggests that substantially all responsive documents have been found; high precision suggests that primarily responsive documents have been found. For an end-to-end review, counting all and only responsive documents that are produced (or withheld on the basis of privilege), as well as responsive documents that are not produced for any reason, a recall of 70-80% is generally considered an acceptable result, and is commensurate with the state-of-the art TAR methods reported in the Richmond Journal study referenced above. A related measure is called *prevalence* or *richness*. Defendant's letter of March 22, 2021 (a copy of which is attached hereto as Exhibit B) refers to this as “*responsiveness rate*”—which is the proportion (*i.e.*, percent) of

responsive documents in the collection to begin with. For example, if there are 100 responsive documents in a collection of 1,000 documents, its prevalence would be 10% (100/1,000).

Prevalence, like recall and precision, can never be known with certainty. At best, the three measures can be estimated using statistical sampling.

11. Defendants' proposed TAR methodologies are inadequate for determining when to stop the TAR process, as well as for conducting validation to ensure the sufficiency of their review process and production set. In particular, as will be explained in greater detail below, their proposals are deficient for four reasons: (i) they rest on an imprecise initial prevalence estimate that is likely to reflect considerable error; (ii) they propose to stop the review based on a metric ("*target recall*") derived from this flawed initial prevalence estimate; (iii) they evaluate how many responsive documents have been missed using an *elusion test* (*i.e.*, a random sample of documents drawn from the set of documents deemed to be non-responsive by *the TAR system only*)—a method which is subject to severe reviewer bias and does not account for responsive documents excluded by search terms or human review. (Elusion also fails to account for non-responsive documents included in the production set either as a result of human reviewer error, or because they are attachments to responsive documents but not in and of themselves responsive); and (iv) they do not include any post-hoc validation procedures beyond their flawed elusion test.

Defendants' Initial Prevalence Estimate and "Target Recall"

12. Defendants' proposed TAR process is based on an initial prevalence estimate derived at the beginning of the review process when reviewers know little about the documents or the matter. Defendants Dow, Covestro, Huntsman, and BASF propose to initiate their review using such a prevalence estimate; they intend to draw a random sample of documents from the

collection to determine how many documents in the *sample* are responsive. They will then use the results obtained from this sample to derive an initial prevalence estimate for the *entire review population* based on extrapolation from the sample to the collection. So, for example, if the initial random sample contains 2,400 documents, and of those, 72 are found to be responsive, the initial prevalence estimate would be 3% (72/2,400), a “realistic responsiveness rate” according to Defendants’ letter of March 22, 2021 (*see Exhibit B at page 3*). In a collection of one million documents, that would mean that Defendants would be able to estimate that there were 30,000 responsive documents to be found ($3\% \times 1,000,000$), but, *it is important to bear in mind that this is merely an estimate and this estimate includes a margin of error*. For the example above, from the prevalence estimate we can only state, with 95% confidence, that the true prevalence is somewhere between 2.35% and 3.76%, and that the collection therefore contains somewhere between 23,500 and 37,600 responsive documents (*see statpages.info/confint.html*, $x = 72$, $N = 2,400$). We do not know, and cannot know, the exact number.

13. Yet several Defendants’ protocols would have the review stop when the number of documents slated for production was 70% of this prevalence estimate—a number that I understand was referenced at a meet and confer between the parties and in correspondence (*see Exhibit B at page 3*)—or 21,000 documents. This might be 89% of all responsive documents (if the true prevalence is 2.35%), but it equally might be **55.9%** of all responsive documents (if the true prevalence is 3.76%). We just don’t know. And, while 89% would be a great result; 55.9% would be unacceptable. Both possibilities are entirely possible because they both fall within the margin of error.

14. Dow and Covestro, on the other hand, propose to stop review “when the responsiveness rate of the documents being reviewed meets or falls below the expected

prevalence for the entire review population.” (See Covestro and Dow TAR Methodology Disclosures of February 19, 2021, attached hereto as Exhibits C and D, respectively). If I understand their statements correctly, this stopping criterion would be equivalent to Plaintiff’s proposal if prevalence is in the range of 5%-10%, *more strict* than Plaintiffs’ proposal if the prevalence is lower, and *more lax* than Plaintiff’s proposal if the prevalence were to be higher. I am not sure why the overall prevalence should play a role in determining the stopping criterion. Assuming the TAR system works, the responsiveness rate in the review sets will initially be quite high (because the whole point of TAR is to identify responsive documents up front)—thus, the review sets will contain far more than 3%—and the number of responsive documents will eventually fall off, regardless of the initial prevalence. If the prevalence is higher than 10%—which is entirely possible if keyword culling is used before TAR—say 30%, it would not be reasonable to terminate the review as long as 29% of the documents being reviewed were relevant, as called for by the Dow and Covestro protocols.

15. Defendants have sown confusion by using margin of error, expressed as a percentage, as in indirect way to specify sample size. I understand that Defendants initially proposed to derive the sample size based on a 95% confidence level and a margin of error of $\pm 5\%$, and in the case of Dow and Covestro, using a standard response distribution of 50%, meaning that they are assuming a prevalence of 50%; Wanhua provides no criteria for their initial prevalence sample. I also understand that Defendants have since agreed to reduce the margin of error to $\pm 2\%$. These criteria determine sample sizes of approximately 385 and 2,400 documents for the “ $\pm 5\%$ ” and “ $\pm 2\%$ ” samples, respectively.

16. Of course, despite what Dow and Covestro assume, the actual prevalence is likely to be considerably less than 50%. As shown above, when the prevalence is on the order of 3%,

the actual margin of error of a prevalence estimate based on a sample size of 2,400 is about **±0.7%**, which sounds like a very small amount. But it is actually quite large when we consider the number of responsive documents to be found: The true value may be between 23,500 and 37,600—*a difference of 60% between these two possibilities (i.e., 37,600 is 60% greater than 23,500)*. For this reason, such an estimate cannot yield a reliable indication of whether or not a production is substantially complete. Defendants should not be permitted to use such a prevalence estimate to determine when to stop the review, because as shown above, this can lead to unpredictable recall (*i.e., an unknown value between 55.9% and 89%*).

17. Aside from the statistical margin of error discussed above, the error of the initial prevalence estimate will likely be substantially greater than the statistical margin of error due to *human error* in reviewing the initial richness sample at the beginning of the review process when reviewers lack knowledge about the documents and are likely to make numerous coding errors, thereby compounding the inaccuracy of the initial prevalence estimate.

18. Defendants' emphasis on the initial prevalence estimate is misplaced. As shown above, their proposals will result in an unreliable stopping criterion for TAR, and will yield a highly imprecise measure of recall. Ironically, Defendants' proposal may also result in disproportionate review effort—they may need to review far more documents than reasonable to achieve their target estimate of recall or “responsiveness rate.” Equally likely, their estimate may result in stopping the review prematurely, achieving the stated target estimate but not actually achieving high recall, in other words, stopping the review effort short, when many more responsive documents could be found with little additional burden. The reasons for this will be explained below.

Defendants' Stopping Criteria

19. Defendants propose to rely on their initial prevalence estimates to determine when to stop the review. Dow and Covestro (as far as I understand what they wrote in their letters) propose to stop the review when the latest batches of documents being reviewed meet or fall below the estimated prevalence for the entire review population. As noted above, this approach may be unreasonably burdensome if prevalence is low, and will terminate the review unreasonably early if prevalence is high. For BASF, the review would stop when its estimate of "sufficient" recall is reached, assuming 70% recall as "sufficient," as I understand was proposed by Defendants at a meet and confer and in correspondence (*see Exhibit B at 3*). What this means is that in a review population of one million documents, if the initial prevalence estimate was 3% (as Defendants claim it will be), and the "recall target" was 70%, BASF would stop the review when 21,000 responsive documents ($70\% \times 30,000$) were identified—whether correctly or not—by the manual reviewers. Huntsman suggests that review will conclude "[w]hen the responsive probability scores and responsiveness rate suggest that we have achieved sufficient recall." And Wanhua describes its stopping criterion as "a target percentage of recall within the TAR Corpus." *See* BASF, Huntsman, and Wanhua TAR Methodology Disclosures of February 19, 2021 (attached hereto as Exhibits E, F, and G, respectively). I understand these to be similar to BASF's criterion, but without specification of the target percent, and subject to the same limitations.

20. Dow and Covestro do not and cannot establish a rational basis for refusing to continue the review until the *marginal precision* (*i.e.*, the responsiveness rate of the last several review batches) drops to 5%-10% precision (meaning that they are finding only 5-10 relevant documents out of 100), which is what Plaintiffs' propose. I have a hard time imagining a

circumstance in which it would be reasonable to arbitrarily discontinue a review process in the midst of achieving a high yield of responsive documents (*i.e.*, over 10%) in the batches that are being reviewed, particularly if the documents are novel and/or more than marginally responsive.

21. To demonstrate further how reliance on the initial prevalence estimate results in an inadequate stopping criteria (as relied on by the BASF and Huntsman proposals), suppose, *arguendo*, that a document collection contains one million documents, of which 30,000 are *actually* responsive (we would not know this number in advance, but let's assume this to be the true number). The true prevalence of this collection would be 3% ($30,000/1,000,000 = 3\%$). If we were to try to estimate this prevalence with a sample of size 2,400 documents, there is 95% probability that the prevalence estimate would be between 2.3% and 3.7%; in other words, that the estimated number of responsive documents would fall somewhere between 23,000 and 37,000 responsive documents.

22. Suppose that Defendants conducted the richness sample and their initial estimate was 23,000 responsive documents (2.3%). If they were to stop the review when the reviewers found 70% of that estimated number (*i.e.*, their “target recall”), that would be when the reviewers had marked 16,100 documents responsive. Assuming *arguendo* that their manual review was infallible—a highly implausible assumption—the actual recall that would be achieved would be **53.7%** ($16,100/30,000 = 53.7\%$), **not 70% as claimed**. Suppose that, equally likely, the richness sample yielded an estimate of 37,000 responsive documents (3.7%), and TAR was terminated at 70% of that estimate—when 25,900 responsive documents had been found by the reviewers. The actual recall would be **86.3%**, **not 70% as claimed** ($25,900/30,000 = 86.3\%$). Under either scenario, Defendants' proposal does not permit the parties to determine the proper stopping criteria and to properly measure recall.

23. Either it is possible to achieve recall of 86.3% with reasonable effort, or it is not.

If it is possible, the first scenario—achieving only 53.7% recall—falls far short of achieving what is reasonably possible. And, if it is not possible to reach 86.3% recall with reasonable effort, the second scenario will result in disproportionate review effort. Either way, Defendants' estimated recall measurement is wrong.

24. Moreover, for reasons that I explain at length in my article with Professor Cormack entitled *Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review,'* in volume 7, issue 1 of *Federal Courts Law Review* (2014) ("Comments paper") at pages 301-10 (a copy of which is attached hereto as Exhibit H)—which would be too lengthy to repeat here—a **prevalence estimate** with a margin of error of $\pm 2\%$ or less is not sufficient to yield a reliable estimate of **recall**. It certainly does not yield an estimate of recall with a margin of error of $\pm 2\%$ or less. In the example set forth above, although the margin of error of the **prevalence estimate** is $\pm 0.7\%$, the margin of error on the **recall estimate** would be $\pm 16\%$. Indeed, Defendants fail to acknowledge how challenging it is to obtain a reasonable margin of error for a valid recall estimate without a disproportionately large sample size; much larger than is being contemplated here. This point is addressed in the *Comments* paper at pages 305-07. Defendants' TAR stopping criterion, based on a false "target recall," is a fool's errand.

25. These numbers also assume perfection by the reviewer(s) of the initial richness sample, as well as by the human reviewers in connection with TAR—something which simply cannot be assumed—and accordingly, the actual margin of error will be far greater than asserted.

Defendants' Elusion Test

26. Once they have reached their designated stopping criteria, Defendants propose to apply an elusion test. This is the only post-hoc validation they propose. Through the elusion

test, Defendants propose to review a random sample of documents that the *TAR tool* deems to be non-responsive, and accordingly, were not subject to any human review. Defendants will review the documents in the sample to determine how many responsive documents are found. So, for example, if the elusion sample consists of 1,000 documents, and 10 of them are found to be responsive, the estimated elusion would be 1% (10/1,000). From this elusion estimate, it is possible to derive a coarse estimate of how many responsive documents were excluded from review *by the TAR process* (*i.e.*, an estimated 1% of the total number of documents deemed by TAR to be non-responsive, but with 95% confidence, we can say only that the true elusion is somewhere between 0.48% and 1.83%). Assuming that 800,000 documents were excluded by TAR, this estimate indicates that between 3,840 and 14,640 relevant documents were missed—a nearly four-fold range of uncertainty. Furthermore, *elusion tells us nothing about how many of the produced documents are responsive, or how many responsive documents were missed by human review, by the application of keyword culling, or by any other means.*

27. Even as a means to estimate just the number of responsive documents missed by TAR, elusion is a poor measure because sampling of only the documents deemed non-responsive by TAR suffers from inevitable *reviewer bias*, whether intentional or not. When reviewers are aware that the documents they are reviewing were previously excluded as non-responsive, they cannot avoid being influenced by this knowledge. Unless the validation reviewers are blind to whether or not documents have been produced, and why, and unless the samples they are reviewing also contain a substantial number of responsive documents, they cannot help but exhibit bias. Even if they are not explicitly informed of the reason the documents were selected for review, their coding decisions will also be influenced by the sparsity of responsive examples. As a result of this bias, *the number of missed documents is improperly counted and the recall*

estimate is improperly estimated. For this reason, if no other, it is imperative that the validation reviewers examine a blended sample of responsive and non-responsive documents representative of the entire collection, regardless of whether they have been identified for production or not (or the reason why), and that such review be *blind*. Otherwise, the validation process itself is confounded and thus faulty.

28. It is worth noting that the only operational difference between Plaintiffs' validation proposal and Defendants' proposed elusion sample—which contains only documents deemed non-responsive *by TAR*—is that Plaintiffs' validation also includes subsamples of documents (“strata”) representing (i) documents produced (or withheld on the basis of privilege), (ii) documents coded non-responsive by human reviewers, and (iii) documents excluded by other culling steps (*i.e.*, keywords). These subsamples provide useful information about *where* in the review process responsive documents were missed and can assist in devising reasonable mitigation strategies where necessary. Defendants can readily derive an estimate of elusion, if they so desire, by confining their attention to the subsample (“stratum”) containing only the documents excluded by TAR. Unlike Defendants' proposals, Plaintiffs validation method does not rely on any purported margins of error; it is a *quality assurance/auditing process* that seeks to determine if and where in the review process systematic mistakes may have been made so they can be corrected, if necessary.

29. As mentioned above, Defendants' elusion test by itself gives no information about *recall*. To demonstrate, if there are one million documents in the collection, and 99% of them are actually non-responsive (*i.e.*, there are 990,000 non-responsive documents), to achieve an elusion of 1%—which sounds really low—the reviewers would only have to mark *every single document* as non-responsive. But while the elusion would appear to be low—“only 1%”—in

reality, *every single responsive document* (*i.e.*, all 1%) would have been missed. I have seen many matters where a low elusion rate has been achieved—and a party has claimed their production was adequate—when the recall was unacceptably low (*i.e.*, less than 50%).

30. Other problems with the Defendants' elusion samples stem from the same concerns that were raised regarding the initial prevalence estimate. Even with an apparently small margin of error—when expressed as a percentage—the range in the absolute number of responsive documents missed will be immense, especially since the elusion rate is expected to be much lower than the 3% initial richness rate that Defendants claim will be the case (this is because prevalence decreases as responsive documents are found).

31. Aside from the elusion sample, Defendants do not propose to do any other validation to determine whether their “target recall” was actually achieved; they simply assume that to be the case.

Defendants' Recall Estimate

32. In addition to the problems set forth above, Defendants also do not propose a method to estimate recall properly—and therefore, a reasonable validation protocol—because (i) they improperly include non-responsive documents miscoded as responsive (*i.e.*, false positives) and non-responsive documents produced along with responsive family members in their “recall” estimate, which will incorrectly inflate that measure; and (ii) they calculate their purported recall only with respect the TAR tool, ignoring documents missed by keywords or by human error, both of which are known to be substantial.

33. Defendants propose to calculate recall using the following formula:

$$\text{Number of Documents Produced} / (\text{Number of Documents Produced} + \text{Elusion Estimate})$$

Putting this in terms of numbers, this means that if Defendants produced 30,000 documents, and the elusion sample estimated that that 10,000 responsive documents had been missed, their estimate of recall would be: $30,000/(30,000+10,000) = 75\%$ recall. Over and above the inherent imprecision and bias of the elusion estimate, this calculation would be incorrect. The number of documents produced is *not* equal to the number of responsive documents produced, because (i) reviewers will incorrectly code some non-responsive documents as responsive, and (ii) some non-responsive documents will be produced along with responsive documents (*e.g.*, attachments). Therefore, 100% precision (*i.e.*, the fact that *only* truly responsive documents were produced) cannot be automatically assumed; *precision must be measured*. Assuming a high level of reviewer precision, say 80%, a valid estimate of the number of produced responsive documents would be 80% of 30,000 = 24,000. And assuming a high level of reviewer recall, also say 80%, we may deduce that 6,000 responsive documents (20% of 30,000) would be incorrectly coded, and thus excluded from the production. So, a valid recall estimate, in this case, would be $24,000/(24,000+6,000+10,000) = 60\%$, *considerably lower than the 75% that would be represented by Defendants*.

34. Defendants' recall estimate is also flawed because the elusion test used to calculate it is biased, as explained above, and because it estimates recall solely on the basis of the TAR portion of the review—only those documents that TAR was applied to and deemed to be non-responsive, and therefore not reviewed by a human. It ignores the significant contribution of reviewer coding error.

35. I understand that the parties have agreed to use search-term pre-culling, which will result in reduced recall *even before the TAR process begins*. An accurate and proper measure of recall evaluates the *totality of the review effort*—including the impact of the keyword

culling, TAR, *and* human error—and therefore, is based on the entire document collection, not just the TAR review set.

36. Validation (and recall calculation) should apply to the end-to-end search and review process, starting with the original collection and ending with the production set, regardless of which population TAR is applied to. That is, *validation must account for all responsive documents potentially excluded by the search and review process*—as well as non-responsive documents included in the production set. Calculating recall only on a smaller body of documents, or a single phase of the review process, necessarily diminishes the validity of the recall measure and misrepresents the actual quality of the production. *See Comments* paper at pages 300-01.

37. To make this multiplicative impact more concrete, imagine a total collection of one million documents of which 100,000 are responsive. Further imagine that 60% of the documents are eliminated through search-term pre-culling, such that 400,000 documents move on to TAR. Let us further assume that 70,000 of the 400,000 documents are responsive. It follows that 30,000 responsive documents would be left behind among the 600,000 documents excluded by the search terms. That is, the recall of the search terms would be 70%. (In my experience, search-term recall would typically be lower than 70%, since search terms often tend to miss more responsive documents than 30%.) If Defendants were to achieve their “target recall” of 70% on the TAR review population of 400,000 documents, they would stop the review when they found 49,000 responsive documents. But the recall of the search and review process would *not* be 70% as suggested; it would actually be 49% (70% x 70%), because of the prior application of the search terms. And even that would be an exaggeration because the human reviewers would miscode a certain number of responsive documents as non-responsive. Let’s

assume the reviewers correctly code 70% of the documents, which is consistent with empirical research and my own experience. Then, the actual recall of the production set would be even lower ($70\% \times 70\% \times 70\% = \underline{\mathbf{34.3\%}}$). That is, only 34,000 of the 100,000 responsive documents would be produced. Even using a broad set of search terms, which might return a greater percentage of responsive documents—let's say 85%—the recall would still drop substantially when calculated for the whole end-to-end process ($85\% \times 70\% \times 70\% = \underline{\mathbf{41.65\%}}$).

38. I would note that the recall values on the order of 70%-80% observed in the Richmond Journal study, which have been cited to characterize an adequate review, reflect the *end-to-end review* of the entire collection as proposed by Plaintiffs, not *just the TAR component*, as proposed by Defendants. At bottom, the way that Defendants propose to estimate recall is highly likely to mask a much lower recall because their method is likely to overestimate the number of responsive documents found and underestimate the number of responsive documents missed.

Plaintiffs' Stopping Criterion and Validation Proposal

39. Plaintiffs' proposed stopping criterion and TAR validation proposal, respectively, are based on empirical research, and a robust and statistically sound methodology that I developed as the special master in the *In re Broiler Chicken Antitrust Litigation*, No. 1:16-cv-08637, 2018 WL 1146371 (N.D. Ill. Jan. 3, 2018) (Order Regarding Search Methodology for Electronically Stored Information) (attached hereto as Exhibit I). They have been successfully implemented in other cases. The review is terminated when the last several hundred documents identified by TAR for human review contain no more than 5%-10% responsive documents, and none of those responsive documents are novel and/or more than marginally responsive. The validation protocol relies on *stratified sampling*. By stratified sampling, I mean that the final

validation sample is comprised of multiple subsamples, including documents produced (or withheld on the basis of privilege), documents excluded by search terms, documents excluded by TAR, and documents excluded by humans, in order to properly estimate the number of relevant documents in each stratum, and to more properly reflect true recall. Recall, calculated in this manner, is estimated by comparing an estimate of the actual number of documents produced (or withheld as privileged)—as opposed to incorrectly *assuming* all produced documents are responsive—to a statistically valid estimate of the number of responsive documents not produced (or withheld as privileged) at all stages of the review process.

40. Plaintiffs' proposed stopping criterion obviates the need for an initial prevalence estimate—removing the burden on Defendants of reviewing 2,400 documents at the outset of the matter—allowing the TAR process to start immediately. Empirical evidence shows that, at the outset, the majority of the high-scoring documents will be responsive, and as the TAR scores diminish, fewer and fewer documents will be responsive. When the proportion of responsive documents drops substantially—to between 5% and 10%—empirical evidence suggests that high recall will have been achieved. *See* Gordon V. Cormack and Maura R. Grossman, *Multi-faceted Recall of Continuous Active Learning for Technology-Assisted Review*, Proceedings of the 38th Annual ACM SIGIR Conference, 763, 765 (2015) (“Our experiments suggest that when a review achieves sustained high precision, and then drops off substantially, one may have confidence that substantially all facets of relevance have been explored. In addition to offering a potentially better prediction of completeness, precision can be readily calculated throughout the review, while recall cannot.”) (attached hereto as Exhibit J).

41. If the proportion of responsive documents continues to be substantial in the review sets (*i.e.*, more than 10%, or more than one in 10 documents is still responsive), it is

obvious that continuing the review will continue to yield responsive documents, for reasonable effort. On the other hand, if the number of responsive documents dwindles to a trickle (*i.e.*, 5%, or only one in 20 documents is still relevant), then further review is likely fruitless and disproportionate. If the responsive documents being identified when the number of responsive documents has dropped are novel and/or more than marginally relevant, this suggests that TAR has had poor coverage. That means there are types of documents in the review set that TAR has not been able to identify for a human relevance determination. In that case, review should continue—or an alternate search strategy should be employed—in order to identify the poorly covered documents and include the responsive ones in the production.

42. Plaintiffs' stopping criteria imposes no additional burden on Defendants, and as demonstrated above, removes the absurd possibility of requiring Defendants to review until more responsive documents are reviewed than exist in the collection, or the equally problematic result of missing a substantial number of responsive documents.

43. Once the review is terminated based on Plaintiffs' proposed stopping criterion, validation begins. Plaintiffs' proposed validation protocol implements stratified sampling, a well-established statistical method to improve the accuracy of estimated recall. The procedure is simple. It requires Defendants to assign a "subject matter expert"—one or more individuals with expertise in the documents and issues in the case—to review a total sample of 5,000 documents made up of documents drawn randomly from four separate subcategories: (i) documents coded responsive by a human; (ii) documents deemed by a human to be nonresponsive; (iii) documents deemed by TAR to be nonresponsive; and (iv) documents excluded from TAR and human review by keywords. The subject matter expert(s) review the documents without knowing the subcollection(s) the documents came from or the prior coding decision(s). When this is done,

Defendants prepare and provide Plaintiffs with a simple table of the results—identifying the number of responsive and non-responsive documents identified in each subcategory. Any responsive documents not previously produced are provided to Plaintiffs. Importantly, the stratified-sample review is *unbiased* because it includes both responsive and non-responsive documents, so the subject matter experts will not have any expectation as to how each document should be coded.

44. The results of the stratified sampling are easily used to calculate recall and precision following simple steps. The number of non-responsive documents in the first subcategory provides an estimate of precision (*i.e.*, the number of non-responsive documents coded responsive, or the number of false positives). The results of the final three subcategories demonstrate elusion (*i.e.*, the number of responsive documents that were missed, or the number of false negatives), whether as a result of human review, TAR, or keywords, respectively. Together, the numbers of false positives and false negatives can be extrapolated to the document population as a whole to calculate a statistically valid estimate of recall.

45. The benefit of the stratified sampling method, which samples the four subcategories, rather than just one—as Defendants propose—is to identify where errors (if any) may have occurred for quality assurance purposes.. For example, if there are a large number of responsive documents in the third subcategory but not in the second, there may be an issue with the TAR system; if there are a large number of responsive documents in the second subcategory but not in the third, then the humans reviewers may have coded inaccurately. With this information, errors can be readily corrected without re-reviewing the entire collection.

46. The stratified sampling method also has the benefit of evaluating the search terms applied to pre-cull the review set—a necessary metric to calculate end-to-end recall. The fourth

subcategory represents the set of documents as to which TAR was not applied because they did not hit on search terms. Typically, to achieve the best results, TAR would be run on *the entire document collection*, not a collection pre-culled using search terms. However, as here, when search terms are being used, sampling to estimate the number of responsive documents in the culled set of documents provides confirmation of the accuracy of the search terms.

47. I understand that Defendants have pointed out that the *Broiler Chicken* validation protocol on which Plaintiffs' validation protocol is based (see Exhibit F), did not include a subsample (or stratum) for documents that were missed by keywords. That is correct. However, a review of the *Broiler Chicken* protocol makes clear why that was the case. In *Broilers* (i) *Plaintiffs got to choose a certain number of key custodians whose documents were not keyword culled at all*, and (ii) *Plaintiffs were given an opportunity to propose a second set of keywords after a substantial initial production had been made by Defendants*. Had that not been the case, it is likely that I would have included the fourth subsample in the validation protocol.

48. Plaintiffs' proposed methodology imposes no meaningful additional burden on Defendants. Defendants' proposal requires sampling of approximately 4,800 documents (approximately 2,400 for the initial prevalence estimate, and approximately 2,400 for the elusion estimate). Plaintiffs' proposal requires sampling of 5,000 documents in total. Moreover, Plaintiffs' proposal does not require exceptional transparency, such as provision of any non-responsive or privileged documents.

49. Plaintiff's proposed methodology provides a sound and unbiased estimate of the number of responsive documents identified by the review process, as well as the number of responsive documents excluded for any reason. These estimates may be combined to yield a coarse overall estimate of recall. A recall estimate in the neighborhood of 70% or more, when

conducted in this fashion, is consistent with a reasonable production, but is not the sole determinant. The adequacy of the production also depends on not finding a significant number of responsive documents that are novel and/or more than marginally responsive during the validation process. In the *In re Broiler Chicken Antitrust Litigation*, where this stratified sampling method was applied, it did not prove to be particularly onerous or an unreasonably difficult standard for producing parties to achieve.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

Executed this 25th day of March, 2021 in Waterloo, Ontario.

s/ Maura R. Grossman, J.D., Ph.D.
Maura R. Grossman, J.D., Ph.D.